Original JAMIA Investigations

Research Paper ■

Evaluation of a Method to Identify and Categorize Section Headers in Clinical Documents

JOSHUA C. DENNY, MD, MS, ANDERSON SPICKARD, III, MD, MS, KEVIN B. JOHNSON, MD, MS, NEERAJA B. PETERSON, MD, MSC, JOSH F. PETERSON, MD, MPH, RANDOLPH A. MILLER, MD

Abstract **Objective:** Clinical notes, typically written in natural language, often contain substructure that divides them into sections, such as "History of Present Illness" or "Family Medical History." The authors designed and evaluated an algorithm ("SecTag") to identify both labeled and unlabeled (implied) note section headers in "history and physical examination" documents ("H&P notes").

Design: The SecTag algorithm uses a combination of natural language processing techniques, word variant recognition with spelling correction, terminology-based rules, and naive Bayesian scoring methods to identify note section headers. Eleven physicians evaluated SecTag's performance on 319 randomly chosen H&P notes.

Measurements: The primary outcomes were the algorithm's recall and precision in identifying all document sections and a predefined list of twenty-nine major sections. A secondary outcome was to evaluate the algorithm's ability to recognize the correct start and end boundaries of identified sections.

Results: The SecTag algorithm identified 16,036 total sections and 7,858 major sections. Physician evaluators classified 15,329 as true positives and identified 160 sections omitted by SecTag. The recall and precision of the SecTag algorithm were 99.0 and 95.6% for all sections, 98.6 and 96.2% for major sections, and 96.6 and 86.8% for unlabeled sections. The algorithm determined the correct starting and ending text boundaries for 94.8% of labeled sections and 85.9% of unlabeled sections.

Conclusions: The SecTag algorithm accurately identified both labeled and unlabeled sections in history and physical documents. This type of algorithm may assist in natural language processing applications, such as clinical decision support systems or competency assessment for medical trainees.

■ **J Am Med Inform Assoc.** 2009;16:806–815. DOI 10.1197/jamia.M3037.

Introduction

Electronic health records comprise a rich source of clinical information, including observations about patients, laboratory and imaging reports, diagnoses, therapies, and longitudinal descriptions of patients' illnesses over time. Most clinical records exist as natural language text narratives that providers compose as they interact with patients and interpret test results. As healthcare providers write, dictate, or electronically enter clinical documents, they typically use conceptual, physical, or electronic templates to divide their narratives into commonly-recognized segments, or *sections*. Clinicians often label the segments with frequently used but nonstandardized terms ("section headers"). For example, history and physical examination (H&P) notes generally contain sections labeled "history of present illness", "past medical history", and "physical examination." Sections can have subsections, such as "cardiovascular exam" within "physical examination" or "substance abuse history" within "social history". We describe development and evaluation of an algorithm, called SecTag, to parse natural-language clinical documents and label these sections. The SecTag algorithm uses an empirically derived document section header terminology.[1] It identifies both explicitly labeled and unlabeled (i.e., implied) sections of clinical narratives.

Background

Motivation for Automated Section Header Identification

Review of seven decades of physical diagnosis textbooks suggests that common section headers found in clinical

Affiliations of the authors: Department of Biomedical Informatics (JCD, AS, KBJ, JFP, RAM), Division of General Internal Medicine and Public Health, Department of Medicine (JCD, AS, NBP, JFP), Department of Pediatrics (KBJ), Vanderbilt University School of Medicine, Nashville, TN; Tennessee Valley Geriatric Research Education Clinical Center (GRECC), Tennessee Valley Healthcare System, Veterans Administration, Nashville, TN (JFP).

This work was supported by the National Library of Medicine grants T15 LM007450 and R01 LM007995 and the National Cancer Institute grant R21 CA116573. The authors appreciate the assistance of Drs. Stéphane Meystre and Peter Haug, who shared their section terminology and section identification algorithms with us at the initiation of our project, and stimulated our thinking about how to expand upon their work.

Correspondence: Joshua C. Denny, MD, MS, Eskind Biomedical Library, Room 442, 2209 Garland Ave, Nashville TN 37232; e-mail: <josh.denny@vanderbilt.edu>.

Received for review: 10/17/08; accepted for publication: 08/03/09.

notes have minimally changed over the last century.[2–9] In medical, nursing, and dental schools most students learn to write clinical notes using these standard section headers.[10–16] Many structured note capture tools use this common organizational structure.[17–21] Within given note types, clinical note sections generally follow a logical sequence, such as the head-to-toe anatomical ordering of physical examination components.[9] Clinical narrative sections also figure into physicians' compensation for clinical encounters, based on the United States. Evaluation and Management Coding ("E/M Coding") system.[22]

Identification of sections within clinical documents provides important context for recognizing and understanding the biomedical concepts they contain. For example, the term "friction rub" (a clinical finding that occurs in multiple anatomical loci) has more specific meaning and clinical implications when one knows if it appeared in the "pulmonary auscultation", "cardiac auscultation", "abdominal examination" or "joint examination" section of a note. Similarly, "past medical history" and "family medical history" sections list names of diseases that, based on context, have different implications regarding who has the disease. Clinical note section tagging enhances the functionality of the following applications: (a) clinical decision support systems that require background information to provide appropriate advice (e.g., does the patient have a history of heart failure?),[23,24] (b) automated problem list generators,[25] (c) systems that support healthcare professionals' education (e.g., has a trainee ever evaluated a patient with pneumonia or reported hearing a diastolic murmur?),[26] (d) clinical documentation tools that convert natural language descriptions into structured representations from a target terminology, such as SNOMED CT®,[17,18,20] and (e) efforts to mine electronic medical records for evidence of completion of quality metrics such as smoking cessation or lifestyle counseling.[27]

Automated detection of clinical document section headers ("section tags") can be problematic. Terms that clinicians use to designate document section tags vary significantly, based on use of acronyms, abbreviations, or synonyms (e.g., the "history of present illness" section might appear as "HPI", "history", or "history of current illness"). A given section tag may be absent from a document altogether—the human reader (or section-tagging tool) must infer the presence of the omitted section header whenever the corresponding section content is present. For example, "40 pack-year history of smoking" implies presence of a "substance use history" section, even when no section label precedes it.

Natural Language Processing and Section Header Identification

To facilitate both human retrieval and automated tools' "understanding" of clinical documents, informatics applications identify target concepts (often from standardized terminologies) within clinical and biomedical texts (e.g., mapping "mad-cow disease" and "bovine spongiform encephalopathy" to the same unique concept identifier). In the 1990s, Carol Friedman and colleagues developed the MedLEE (Medical Language Extraction and Encoding) system, arguably one of the most comprehensive existing clinical natural language processing (NLP) systems. The MedLEE system recognizes concepts within many clinical document types including discharge summaries,[28,29] radiograph reports,[30] mammograms,[31] and pathology reports.[32] It includes a document preprocessor that recognizes some common clinical section tags (e.g., "history of present illness"),[33] though no formal evaluation of the segmenter's accuracy has been published.

Several clinically-oriented NLP projects, including, among others, SAPHIRE,[34] MetaMap,[35] the KnowledgeMap Concept Identifier ("KnowledgeMap"),[36,37] a system developed by Nadkarni and colleagues,[38] the Mayo Vocabulary Processor,[39,40] and systems developed by Chapman and colleagues,[41–43] took unique approaches to identifying UMLS (or other standard vocabulary) concepts within clinical texts. While many of these systems index clinical documents at the sentence or noun-phrase level, they often do not explicitly segment clinical notes by section. Thus, such systems may not distinguish the different meanings of "congestive heart failure" when the phrase appears in the family history, past medical history, or the assessment sections of clinical notes. However, some systems correctly classify concepts when a sentence within a note contains clues such as, "there is a family history of colon cancer."

Meystre and Haug created a NLP system to generate problem lists by processing clinical documents.[25,33] Their first step used a document parser that matched document strings to a list of known headers to identify common sections within clinical documents. Their system assigned all text from the beginning of one section header to the start of the next recognized section header to the former section header. While exemplary, their algorithm misclassified unknown or unlabeled sections.

This paper describes a general algorithm, SecTag, for identifying section headers and delimiting the content associated with those sections. Authors evaluated SecTag's performance on a corpus of H&P notes. The algorithm can serve as a preprocessor for other NLP applications.

Methods

Methods—SecTag System Overview

Using a locally-developed section header terminology, SecTag identifies section tags within natural language clinical documents.[1] The tagger identifies explicitly labeled section headers and deduces presence of implicitly labeled (i.e., missing) section headers. The SecTag algorithm produces output in two formats, native XML with identified section headers annotated within the note's original text, or as an HL7 CDA-compliant XML document with identified section labels (see online Appendix 1, available at http://www.jamia.org, for hypothetical notes with corresponding SecTag CDA output). The Vanderbilt University Medical Center Institutional Review Board approved the SecTag development process and evaluation.

Methods—H&P Corpus for SecTag Development and Evaluation

To develop the SecTag algorithm and the terminology, we divided a corpus of deidentified electronic H&P documents into a training set and an evaluation set.[1] Vanderbilt's EMR system captures all inpatient and outpatient notes since 2001, and a large portion of notes since the early 1990s. To provide H&P notes for this study, an automated program randomly selected 25,000 notes within the Vanderbilt EMR

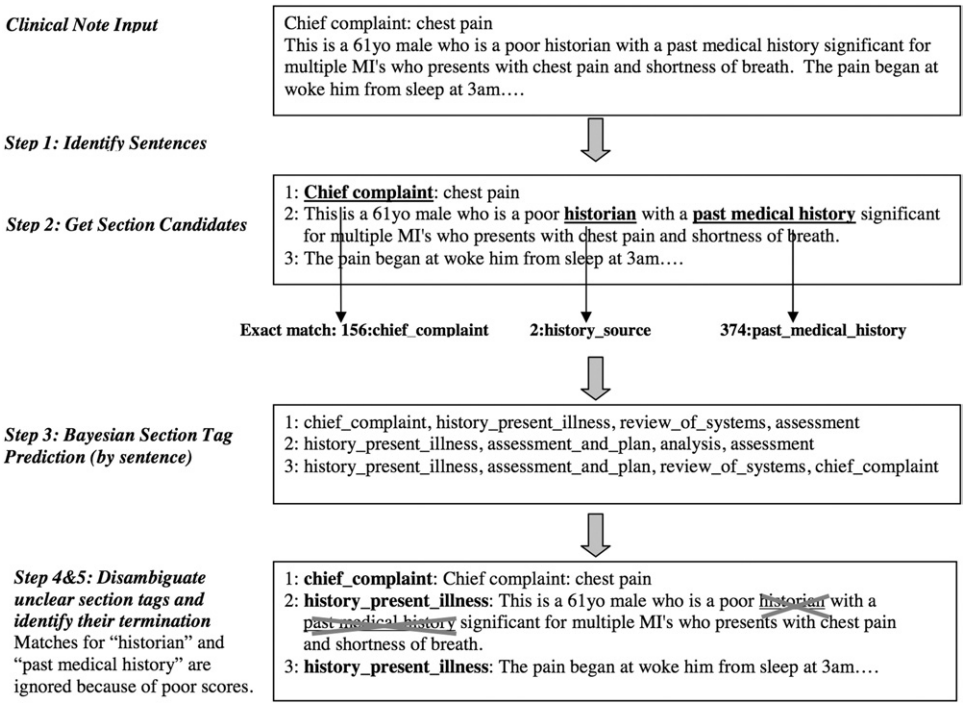

Figure 1. SecTag Processing of a Section of Text.
The "Chief complaint" tag is an exact match since it starts a line, matches a string exactly, and ends in a colon. The Bayesian score of the next sentence highly favors the "history_present_illness" because it follows the chief complaint, occurs toward the beginning of the document, and contains words common for this section. Thus, "historian" and "past_medical_history" are ignored as possible tags and the section is labeled as "history_present_illness."

system with titles containing "H&P", "admission", or "history." These notes were deidentified by removing the 18 HIPAA safe-harbor categories of information using DE-ID®, a commercially-available software package (University of Pittsburgh Medical Center, Pittsburgh, PA), and other pre- and post-processing refinements.[44] Healthcare providers had originally created the H&P notes by manual entry (typing them into the EMR freehand or via a template) or through dictating notes that were then transcribed into the EMR system. We manually reviewed this set of 25,000 notes and retained only those that were actual H&Ps (n = 10,767). This set was divided into a training set (n = 9,567) used for development of the terminology and SecTag, and an evaluation set (n = 1,200, of which 540 were actually used in the study).

Methods—Section Header Terminology

We previously described development of the section header terminology, which included all relevant section tags for H&P documents.[1] Terminology development derived candidate header terms from LOINC®,[45] the QMR findings hierarchy,[46–48] history and physical examination textbooks,[2–9] the section header list identified by Meystre and Haug,[33] extensive automated and manual review of the H&Ps from our training set, and from review of general and subspecialty H&P templates (n = 83) that exist within Vanderbilt's web-based clinical note writing application.[17] The SecTag terminology data model followed principles established by Cimino[49] and others.[50] Similar to the UMLS,[51] the SecTag header terminology included a concept-oriented structure with polyheirarchical parent–child relationships (e.g., "jugular venous pulse exam" was a child of both the "cardiovascular exam" and the "neck exam"). The SecTag section header terminology contained 1,109 concepts with 4,332 terms. It contains mapping to existing LOINC® and the E/M coding system terms (where possible). The SecTag terminology is available free-of-charge by contacting author JD.

Methods—SecTag Algorithm

The SecTag algorithm sequentially processes documents in five steps (per Figure 1): (a) identify sentence boundaries and elements of lists (e.g., "1. Congestive heart failure"); (b) identify all candidate section headers using lexical tools, spelling correction, and NLP techniques; (c) calculate the Bayesian probabilities that each sentence belongs to any given section; (d) disambiguate unclear section headers, using the Bayesian probabilities; and (e) identify the end (terminal boundary) of each section. We used the training set of H&P notes to develop and iteratively improve the algorithm. The SecTag algorithm is described briefly here; a more detailed description is provided in Appendix 2, available online at http://www.jamia.org.

To disambiguate unclear section headers and to predict where unlabeled sections occur, SecTag employs both statistical and hierarchical models of section headers. The statistical model, derived from automated processing of the training set, contains the prior probability of each possible section header occurring in a document, the probability of each section header occurring in each location in the document, and the probabilistic sequential order for section headers appearing in documents (e.g., "chief complaint" occurs before "history of present illness"). The SecTag algorithm uses the header terminology's parent–child hierarchical relationships (e.g., "substance abuse history" is part of "social history") for rule-based inferences (e.g., the label "mother" likely represents "mother's medical history" if found within the "family medical history" section) and to calculate the path length between two section concepts (i.e., the number of section concepts [nodes] traversed between two section headers in the terminology).

Identifying Sentences and Lists

Clinical notes are often not well-formatted or well-structured, especially when created by busy practitioners through dictation without a template. To "normalize" clinical documents, SecTag first identifies individual sentences and lists within the document. Sample H&P notes used in this study were often word-wrapped through the insertion of newline characters within sentences, so SecTag uses a variant of the KnowledgeMap sentence identifier, to combine sentence fragments separated by newline characters.[36] To aid in section boundary prediction, SecTag identifies paragraph boundaries and enumerated lists of words, phrases, or sentences (e.g., "1. Congestive heart failure . . . . 2. Diabetes . . . ."). A group of sentences within a list likely belongs to the same section or subsection (e.g., if list *X* was immediately preceded by "Assessment and Plan", all its components are likely "assessment_and_plan" elements).

Identifying Candidate Section Headers

The SecTag algorithm next processes the document, sentence by sentence, to find all possible section headers. First, SecTag attempts to identify all explicitly labeled section tags by searching for strings that begin sentences, consist of only capital letters, or end in a dash, colon, or period (because manual review found these delimiters common in section headers in both dictated and electronically generated notes). The system also identifies within-sentence strings that match section header terms, such as "RR: 12, Temp: 37", which match the "respiratory rate" and "temperature" subheaders of the "vital signs" within "physical examination". To further improve section tag recognition, the system employs three techniques to identify document header strings that do not exactly correspond to entries in the SecTag header terminology: (1) recognize derivational word variants and synonyms that map to known terms;[36] (2) correct spelling errors using the open-source spell checker Aspell;[52,53] and (3) remove common "stop words" (such as prepositions and determiners) and certain modifiers such as possessive words, numbers (written as either a word or a number), anatomical references (e.g., "right", "superior", or "bilateral"), and other common adjectives (e.g., "recent", "other"). Each candidate section header within the SecTag lexicon receives a score based on its similarity to each phrase in the clinical document; strings with direct matches receive higher scores than those matched via transformations, such as spelling correction or word variant generation.

The SecTag algorithm also uses syntactic and semantic methods (based on the SecTag terminology) to increase recognition of implied section headers. It applies NLP algorithms to detect section tags that occur within a sentence by examining noun phrases, such as "chief complaint" from "Mr. X is here for a chief complaint of dyspnea." Using the section relationships present in the section terminology, SecTag can find common ancestral section headers when multiple valid section header candidates exist (Figure 2).

When processing a document, SecTag keeps track of all instantiated (i.e., already matched) sections and whether a given section is "open" (or "active"—meaning that new sentences are still being assigned to it). Sections can be nested such that multiple sections are open simultaneously, e.g., both "physical exam" and "cardiovascular exam" can be open.

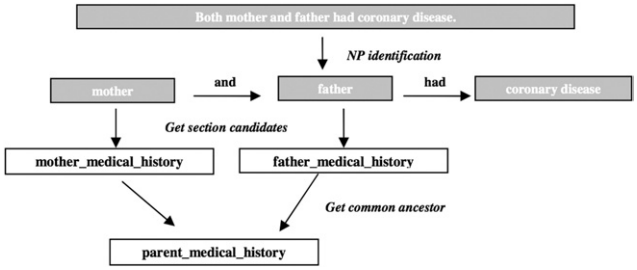

Figure 2. Extraction of Possible Section Candidates Using Noun-Phrase Processing and Common Ancestors. In this example, possible section candidates for "mother" and "father" would be extracted and kept as possibilities until disambiguation, along with the new section "parent_medical_history," added because it is an ancestor relating both "mother_medical_history" and "father_medical_history."

Bayesian Section Tag Prediction

The SecTag algorithm uses naive Bayes scoring to predict unlabeled (implied) section headers, to choose among multiple possible section header candidates for a given text segment, and to discard candidate section headers with low scores. The algorithm calculates Bayesian probability, sentence-by-sentence, for all of a document's section header candidates $Section_i$ based on probabilities of words occurring in each section in the training set ($P(Section_i|\overline{words})$). The training set also determined the probability of any section following a previously encountered exact-matched section in the document ($P(Section_i|priorSections)$), where *priorSections* represents the preceding exact-matched sections in the document. The complete SecTag calculation is

$$P(Section_i \mid \overline{words}, \overline{priorSection}) = P(Section_i) \times \prod_{k=1}^{m} P(Section_i \mid priorSection_k) \times \prod_{j=1}^{|\overline{words}|} \frac{P(word_j \mid Section_i)}{P(word_j)}$$

Training SecTag did not require manual curation—see Appendix 2 for details.

Disambiguating Unclear Section Headers

Many H&P strings potentially map to multiple section header concepts; for example, "cardiovascular" can refer to "cardiovascular_exam", "cardiovascular_plan", "cardiovascular_system_review", etc. Other section headers may be implied, such as when "2/6 holosystolic murmur" connotes "cardiac examination" within "physical examination." The SecTag algorithm calculates a score for all ambiguous sections labels and for all unlabeled sections predicted by NLP or semantic rules. The algorithm uses this score to either select the best scoring section candidate or discard the candidate as a "poor match" (Figure 1). The score for each section involves three major components: (1) the similarity of the clinical note phrase to the candidate SecTag term string, (2) the Bayesian score (calculated as above), and (3) the distance of the candidate section header to other nearby note section headers, calculated as the path length between the concepts in the SecTag terminology. The SecTag algorithm then instantiates the best-ranking section header as "active" (unless it is considered a poor match due to a low Bayesian score or is eliminated through a series of rules—Appendix 2, Section D). For

example, one rule excludes any predicted section candidate that occurs elsewhere in the document as an exact-matched header. Appendix 2 describes the disambiguation methods in detail.

Identifying Section Termination

As SecTag sequentially processes the document, it assigns each encountered sentence to all currently active sections. To determine when to terminate (close) an active section, SecTag employs Bayesian scores and a series of rules. Termination rules are specific to the type of section currently active; for example, short sections such as "chief complaint" are terminated when SecTag encounters a blank line between paragraphs, while longer sections like "history of present illness" can cross paragraph boundaries. The algorithm uses "knowledge" of document formatting to keep an active section open or to close it (e.g., it will not close in the middle of a numbered list, but will do so at its end if subsequent markings suggest a new section header). If a newly encountered word or phrase has a sufficiently high Bayesian score to indicate that it is a new section header that is not a child of any active section header, SecTag may close some or all the open headers. Conversely, to prevent assignment of a sentence to a poorly matching section, an active section header may be dropped from consideration (terminated) if its Bayesian score falls to a sufficiently low-level. Thus, some sentences may not be assigned to any section if the best ranking section receives a poor score.

Methods—SecTag Evaluation

We recruited and paid a small fee to eleven board-certified or board-eligible Vanderbilt-affiliated physician-evaluators. Four were board-certified internal medicine physicians, one a board-certified pediatric intensivist, one a board-certified family medicine practitioner, and four board-eligible internal medicine physicians. None were familiar with SecTag or its section header terminology before recruitment and participation. Each evaluator was asked to review and rate 60 H&Ps from the study evaluation set via a custom Web interface.

The primary objective of the evaluation study was to measure how well the SecTag algorithm could detect the presence (explicitly labeled or implied) of "major" section headers in H&P notes. Prior to the evaluation, project clinicians created, by consensus, a list of twenty-nine common and important major section headers based on review of training set H&Ps (see Table 1 for a complete list). We asked the physician evaluators identify all major sections in each of the 60 documents presented to them. Physicians reviewed each document via a Web interface, which displayed the original deidentified H&P document in one window, side-by-side with a window that showed the SecTag marked-up version of the document, indicating where the SecTag algorithm had identified section headers and section boundaries. Evaluators used the web interface tool to mark each SecTag-identified section header as "correct" or "incorrect." Evaluators also

Table 1 ■ Recall and Precision for Each Major Section

Section Name	*n*	Number Labeled (%)	Recall (95% CI)	Precision (95% CI)
Chief complaint	283	280 (99%)	100% (98–100)	100% (98–100)
History present illness	353	281 (80%)	99% (98–100)	93% (90–96)
Past medical history	296	282 (95%)	99% (98–100)	99% (97–100)
Family medical history	255	250 (98%)	100% (99–100)	98% (95–99)
Parent medical history	192	7 (4%)	95% (90–0.98)	90% (84–94)
Sibling medical history	38	3 (8%)	87% (71–95)	97% (84–100)
Child medical history	3	1 (33%)	100% (29–100)	100% (29–100)
Health maintenance	92	91 (99%)	100% (96–100)	100% (96–100)
Personal and social history	267	252 (94%)	100% (98–100)	99% (96–100)
Substance use history	254	138 (54%)	94% (91–97)	98% (96–100)
Medications	282	254 (90%)	100% (98–100)	99% (96–100)
Allergies and adverse reactions	254	249 (98%)	100% (99–100)	100% (99–100)
Review of systems	462	437 (95%)	100% (98–100)	95% (92–97)
Physical examination	336	307 (91%)	100% (99–100)	99% (97–100)
Vital signs	333	201 (60%)	99% (97–100)	92% (89–95)
General examination	268	211 (79%)	99% (97–100)	100% (98–100)
Dermatologic examination	216	172 (80%)	99% (97–100)	95% (92–98)
Lymph nodes/heme examination	142	131 (92%)	99% (96–100)	99% (96–100)
HEENT examination	767	595 (78%)	98% (97–100)	98% (96–99)
Cardiovascular examination	293	235 (80%)	100% (98–100)	98% (96–99)
Gastrointestinal examination	295	225 (76%)	99% (98–100)	97% (94–98)
Chest examination	374	291 (78%)	99% (97–100)	98% (96–99)
Genitourinary examination	138	116 (84%)	99% (96–100)	94% (89–97)
Neurological examination	320	244 (76%)	97% (94–100)	95% (91–97)
Psychological examination	67	62 (93%)	100% (95–100)	99% (92–100)
Musculoskeletal examination	82	51 (62%)	95% (87–99)	92% (84–97)
Extremity examination	314	210 (67%)	97% (95–99)	94% (90–96)
Laboratory, imaging, and pathology results	393	246 (63%)	98% (96–99)	88% (84–91)
Analysis, assessment, and plan	600	503 (84%)	98% (96–99)	96% (94–97)
Total	7969	6325 (79%)	98.6% (98.3–98.8)	96.2% (95.8–96.6)

"Number labeled" refers to the number of sections that had an author-placed label in the document. HEENT = Head, Eyes, Ears, Nose, Throat.

determined and annotated the accuracy of SecTag-identified section boundaries (whether the algorithm had correctly marked the start and end of the text belonging to each section). Using the web-based tool, the evaluators indicated where section headers (unlabeled major sections or author-labeled sections) should have occurred, but were not detected by the SecTag algorithm. To assist the evaluators with their ratings, the physicians were given a reference list of all possible terminology section headers. This list was organized hierarchically (according to the canonical relationships in the SecTag section header terminology) and visually weighted such that more common sections were in larger fonts (based on their frequency in the training documents).

To improve concordance among evaluators, project members provided a training example, followed by a session in which all physician-evaluators scored the same five initial H&Ps. Their ratings were compared, and we established that adequate agreement existed among evaluators. Next, we configured the evaluation sequence such that every eighth document that each evaluator subsequently reviewed came from a common shared document pool. This allowed calculation of inter-rater agreement.

We instructed the evaluators to examine the clinical note's original labels (when present) as a guide to the document author's intended section header when evaluating SecTag's choice of a section header. For instance, if the document included the section label "Head, Eyes, Ears, Nose, Throat" but that section contained only an ear exam, the evaluators were instructed to mark SecTag's annotation as correct if it contained either the broader ("HEENT exam") or more specific ("ear exam") section concept. Likewise, "He has a 40 pack-year history of smoking" could be accurately tagged as the "tobacco use" segment of the "social history" even if the author had placed it in the history of present illness section.

SecTag Evaluation Measurements

The primary evaluation outcomes, based on physician-evaluators' ratings as the gold standard, were SecTag's recall and precision for all major sections listed in Table 1 (whether explicitly labeled in the document or implied), and the recall and precision for all sections (whether major or not) that were identified by the system. We classified SecTag's correctly identified section headers (i.e., where physicians agreed with SecTag) as true positives (TP), and instances where physicians did not agree with SecTag labeling as false positives (FP). Sections labeled by physician-evaluators but not labeled by SecTag comprised "omitted sections (OS)". The study defined the following: (1) OVERALL RECALL: the number of TP divided by the total number of sections (TP + OS); (2) OVERALL PRECISION: the ratio of SecTag's correctly labeled section count to its total number of proposed section tags [TP/(TP + FP)].[3] MAJOR SECTION RECALL: [MTP/(MTP + OMS)], where MTP is the number of major section true positives, OMS is the number of omitted major sections, and (MTP + OMS) represents the total number of gold standard major sections in the document; (4) MAJOR SECTION PRECISION: [MTP/(MTP + MFP)], where MFP is the number of major section false positives.

SecTag Evaluation Statistical Analyses and Sample Size Calculation

Prior to the study, we determined that analysis with 90% power at a precision of 0.9 within a 95% confidence interval of 0.05 would require evaluators to label 471 section headers (or about 26 typical documents from our training set). We chose a much larger set to determine accurate confidence intervals for major sections, which might not be present in every document.

We calculated inter-rater agreement via Cohen's Kappa. We used the Wilcoxon's rank sum test to compare nonparametric data (expressed as median and interquartile range), the Student's t test for parametric data (expressed as mean ± standard deviation), and the χ^2 statistic for categorical data. Confidence intervals were calculated using the binomial exact method. All statistical analyses were performed with Stata, version 9.2 (StataCorp LP, College Station, TX).

Results

SecTag Recall and Precision

The physician-evaluators scored 319 unique H&P documents. Of these, 66 were authored by attending physicians and 252 by house staff; only one note was written by a medical student. Reviewers classified 88% of the notes as full H&Ps, 6% as attending attestations, and 6% as brief admission notes. Forty-four percent of the notes were general medicine service H&Ps, 43% were from nonsurgical subspecialties, 10% were surgical service H&Ps, and 4% were intensive care unit admission notes. Sixty-three (20%) of the H&Ps had been entered using electronic templates; the remaining 80% were dictated and transcribed into the EMR. The SecTag algorithm identified 16,036 sections from the evaluation set (median 52 sections per document); 7,858 (49%) were major sections, such as chief complaint, physical exam, and assessment and plan (Table 1). The 16,036 identified sections contained 355 different header concepts from the SecTag terminology. The training and evaluation sets were similar in numbers of words, numbers of sections, and numbers of labeled sections per H&P document.

Of the 16,196 total sections identified, physician reviewers classified 15,329 sections as true positives, 707 as false positives, and 160 as omitted sections (i.e., false negatives). The SecTag algorithm recall and precision were 99.0 and 95.6% for all section concepts and 98.6 and 96.2% for major section headers (Table 2). The algorithm more effectively identified labeled sections than unlabeled ones (recall 99.8 vs. 96.6%, $p < 0.001$). Table 1 shows the recall and precision for each major section type. The algorithm effectively identified labeled and implicit major section headers with recall 98.6% (range 87–100%) and precision 96.2% (range 90–100%). Document authors often failed to provide section labels for substance abuse history, vital signs, laboratory and radiology results, and first-degree relative family medical history (only 5% were labeled). The SecTag algorithm identified these unlabeled sections primarily with noun phrase processing and Bayesian prediction. Although differences were small, recall was slightly better for nonmajor sections (recall 99.3 vs. 98.6%, $p < 0.001$) and precision slightly better for major sections (95.0 vs. 96.2%, $p < 0.001$). When processing a document, the SecTag algorithm often identified major sections by finding more specific subsections and deducing

Table 2 ■ Recall and Precision for SecTag Section Header Tagging

	Label in Document	No Label in Document	Total
All Sections			
Gold standard	11476	4720	16196
SecTag identified (TP+FP)	11456	4580	16036
Number tagged correctly (TP)	11353	3976	15329
Number tagged incorrectly (FP)	103	604	707
Number where SecTag omitted correct tag (FN)	20	140	160
Recall	99.8% (99.7–99.9)	96.6% (96.0–97.2)	99.0% (98.8–99.1)
Precision	99.1% (98.9–99.3)	86.8% (85.8–87.8)	95.6% (95.3–95.9)
Major sections only			
Gold standard	6325	1644	7969
SecTag identified (TP+FP)	6321	1537	7858
Number tagged correctly (TP)	6250	1310	7560
Number tagged incorrectly (FP)	71	227	298
Number where SecTag omitted correct tag (FN)	4	107	111
Recall	99.9% (99.9–1.00)	92.4% (91.1–93.8)	98.6% (98.3–98.8)
Precision	98.9% (98.6–99.1)	85.2% (83.5–87.0)	96.2% (95.8–96.6)

TP = True positive; FP = False positive; FN = false negative. The gold standard was composed of physician review of SecTag-identified sections and manual identification of sections not labeled by SecTag.

presence of the parent major heading, especially when major sections were unlabeled. For example, SecTag might identify "ophthalmic exam" instead of "HEENT exam" for a focused eye examination, or separately tag the "cranial nerve exam" section of a "neurologic exam"; in these cases, tags were counted as a major sections since the terminology relates them as children of major section concepts. Thus, a search for "neurologic exam" could easily retrieve a segment labeled "cranial nerve exam" regardless of the presence of specific "neurologic exam" section tag.

Evaluators identified 160 (1.0%) section tags that SecTag failed to identify (i.e., for which SecTag did not generate labels). These were either major sections (69%) or document-labeled important nonmajor sections. Most (96%) of SecTag's omitted major sections were unlabeled in the document. Of all missing sections with document labels, all but 13 were present in the section terminology, e.g., "nose and ear exam."[1]

The SecTag algorithm correctly identified the starting and ending boundaries for 92.7% (14,203/15,328) of the correctly labeled sections (Table 3). The system better predicted the boundaries for labeled sections than unlabeled ones ($p < 0.001$). The most common error was an ending error (5.5% [846/15,328] of all correctly tagged sections, 75% of all boundary errors), meaning SecTag either predicted the ending of the section (failed to include relevant content) or too late (included content that did not belong to that section). An analysis of 112 randomly selected incorrectly placed boundaries revealed that 56% of the boundary errors excluded relevant content (i.e., the section was terminated too early) from the section while 31% included too much information. About 15% of the boundary errors were due to nonclinical content in the section, such as outline headers, medical record numbers, or page numbers.

Table 3 ■ Accuracy of Section Boundary Detection for Correctly Labeled Sections

Section Boundaries	Label in Document (%)	No Label in Document (%)	Total (%)
Correct	10983 (96.7%)	3221 (81.0%)	14204 (92.7%)
Incorrect start	20 (0.2%)	197 (5.0%)	217 (1.4%)
Incorrect end	344 (3.0%)	502 (12.6%)	846 (5.5%)
Incorrect start and end	6 (0.05%)	56 (1.4%)	62 (0.4%)
Total			15329

The inter-rater reliability on accuracy between all reviewers was good (Kappa = 0.70, $p < 0.0001$). Each evaluator reviewed an average of 36 H&Ps of the 60 assigned, and 11 documents were scored by multiple evaluators. Inter-rater reliability for placement of section boundaries was lower than for section header identification (Kappa = 0.49, $p < 0.0001$).

Precision of Section Identification Techniques

Table 4 shows the precision of each of SecTag's component algorithms. Spelling correction was the worst performing algorithm, with a correct section prediction occurring with a precision of 62%. Correct spell-correction mediated matches included both multi-word and single-word matches (e.g., "cheif complant", "labarotory"); all incorrect spell-correc-

Table 4 ■ Precision of SecTag Component Methods to Identify Sections

Method	Count (%)	Number Correct	Precision (95% CI)
Labeled Sections			
Exact or normalized match	11221 (70.0%)	11123	99% (98.9–99.3)
Variant generation	130 (0.8%)	110	85% (77–90)
Unlabeled sections			
Bayesian prediction	1867 (11.6%)	1503	81% (79–82)
Next-section rules	29 (0.2%)	27	93% (77–92)
NLP	2112 (13.2%)	1939	92% (91–93)
Both labeled and unlabeled sections			
Spelling correction	53 (0.3%)	33	62% (48–75)
Labels within a sentence	471 (2.9%)	444	94% (92–96)
Modifier removal	153 (1.0%)	150	98% (94–100)
Totals	16036	15329	96% (95, 96)

CI = confidence interval; NLP = natural language processing.

tion mediated matches were single-word matches from document text. Eight spell-correction mediated errors were due to incorrectly disambiguating between the possible sections for an accurate spelling correction (e.g., "ucolor", meaning urine color, became "skin color" since urine color was not a defined section), four were abbreviation/acronyms not present in the terminology (including a deidentified person's initials), and six were the result of SecTag choosing a wrong spelling correction.

SecTag generated 1,664 possible sections headers for which it considered the "best" candidate section header a poor match and thus discarded it. A poor Bayesian score was the most frequent reason to discard a possible section label (58% of discarded sections). The authors evaluated 20 random notes to determine if the poor matches were appropriately discarded or not. Manual review suggested that 93% of the poorly matched section headers were appropriately discarded; 7% could have been instantiated as a section rather than discarded, but none were major section headers.

Discussion

The current study is one of the first large-scale efforts to formally evaluate a clinical note section header identification algorithm. To identify section labels in documents and predict where unlabeled sections belong, SecTag effectively combines NLP methods, concept matching approaches involving variant recognition, and scoring algorithms that include a naive Bayesian classifier. Using a large sample of general H&P documents from the EMR system of a single institution, we found that SecTag accurately identified the great majority of common section headers and boundaries. The algorithm employed a standardized section header terminology, which represented H&P section header concepts well, at least within the test institution.[1]

Accurate section identification is a key first step toward greater automated or semiautomated clinical note processing. In future applications, section identification might be coupled with other NLP tools to improve decision support programs or to enhance clinical research. For example, a decision support system, operating on contextual understanding of concepts within a note, could suggest that a patient with a family history of colon cancer in a first degree relative or with a past personal history of ulcerative colitis should undergo early and more frequent colorectal cancer surveillance.[54]

Future tools might use the SecTag algorithm to improve competency assessment regarding trainee's clinical education. Such a tool might scan medical students' or resident physicians' notes to evaluate the completeness of their workups for patients with a specific condition. The tool might examine all trainee-generated notes on patients with back pain, to assess whether trainees elicited certain key faculty-designated history and physical examination elements (e.g., the presence of saddle anesthesia, incontinence, weakness, or weight loss). The tool could detect whether trainees discussed appropriate "red flag" diagnoses in the assessment and plan section. Toward this end, students and faculty at our institution are using a "learning portfolio" that couples SecTag with the KnowledgeMap concept identifier to track students' clinical experience.[55]

The ability to assign a block of text within a clinical document to a predefined concept within a section terminology hierarchy may improve concept identification, much in the same way as the Linguistic String Project found improved understanding by programming specific sublanguage grammars.[56] For example, the acronym "BS" in the respiratory/chest examination section of the physical examination likely means "breath sounds" but means "bowel sounds" in the abdominal examination section, and possibly "blood sugar" (glucose) in the laboratory results section; likewise, such descriptors as "normal", "non-tender", or "not enlarged" may occur within many sections but indicate distinctly different clinical meanings and evoke different differential diagnoses based on context. For example, when parsing a "vital signs" section, a program might treat any floating point number between 35.0 and 40.0 as a Celsius temperature measurement, and a percentage between 60 and 100% as an oxygen saturation determination, especially if preceded by a segment labeled as a respiratory rate.

The current study's failure analysis revealed several venues for potentially improving SecTag's performance. Some SecTag errors were secondary to suppression of note text by DE-ID software "bugs". The latter program treated certain disease eponyms as being a person's name and "de-identified" them into uninterpretable terms; it sometimes also removed acronyms that were section labels (e.g., "GI"), ostensibly because the acronym matched the initials of a person (e.g., the patient, a physician, or a nurse). A second cause of failure was the spelling correction algorithm, which performed suboptimally. Because the Aspell algorithm used in the study did not support words containing numbers (e.g., the "S4" heart sound), SecTag performance degraded when these words were inappropriately "spell-corrected" to words without numbers. We have since adjusted the algorithm to omit such words from spell-correction. A few common medical words were missing from the spell-check vocabulary; these were added after completion of the study. Finally, dictated documents often contained various forms of deidentified patient names, medical record numbers, and page numbers as new page headers—these often caused errors in section tagging. Imperfect sentence parsing also caused some errors in section boundary detection.

The Bayesian algorithm predicted correct placement of unlabeled sections with an accuracy of 81% for 631 different possible sections. This was not as accurate as other methods SecTag uses for identifying unlabeled section headers. However, the Bayesian score was critical in discarding erroneous candidate sections. A possible cause of error in the Bayesian prediction was an imperfect gold standard, derived automatically from iterative tagging of the training corpora. While the naive Bayes approach performed acceptably, more sophisticated algorithms, such as support vector machines, may perform better.

Limitations of the Current Study

There were several limitations of this study. The study used clinical notes from a single medical center; formatting, styles, typical content, and section header names may be different in other settings. We attempted to mitigate this bias by deriving the SecTag header terminology from external, nationally available sources such as common textbooks and existing standard vocabularies such as LOINC® and QMR.

Second, the prior probabilities for Bayesian scoring were derived from automatically tagged documents (instead of manually tagged documents). While this allowed quick derivation of a large tagged corpus accurate for most sections headers, a manually tagged corpus, would likely be more accurate, were it feasible to create one. Since the nonprobabilistic tagging performed better on major sections than subsections, SecTag is biased towards predicting parent concepts. We only evaluated the SecTag application on H&P documents; other document types (e.g., discharge summaries or operative reports) are likely to require additional training for optimum performance.

The performance of SecTag in identifying subsection tags may have been overestimated in some areas due to physicians' original use of electronic templates in generating notes (20% of this corpus). Furthermore, evaluators were specifically told that matching specific laboratory and radiology subheaders was not a goal of this study. Labeling laboratory and radiology subheaders is less important because, first, these categories are already well represented in existing terminologies such as LOINC®, SNOMED CT®, and QMR, and, second, most EMRs provide laboratory and radiology results in structured and labeled formats. Physician-evaluators scored the SecTag output instead of creating a *de-novo* gold standard, which may bias them in favor of algorithm. Finally, evaluators may have been subject to an information bias in rating nonmajor section headers since they were instructed to concentrate on identifying major section headers.

Conclusions

The current study provides one of the first formal evaluations of a clinical note section header tagging algorithm. The SecTag algorithm accurately identified both labeled and unlabeled sections in randomly chosen H&P documents. The SecTag terminology contained appropriate matches, with very few exceptions, for the section header labels actually used by clinicians during patient care in their H&P notes. Additional research should evaluate and extend the current SecTag terminology and algorithms for other documents types. In the future, the SecTag algorithm could be coupled with a robust concept identification system or a NLP system to provide a better contextual basis for disambiguating clinical terms embedded within source clinical documents.

References ■

1. Denny JC, Miller RA, Johnson KB, Spickard A, III. Development and Evaluation of a Clinical Note Section Header Terminology. AMIA Annu Symp Proc 2008 Nov 6:156–60.
2. Norris GW, Landis HRM, Krumbhaar EB, Montgomery CM. Diseases of the Chest and the Principles of Physical Diagnosis, 5th edn, Philadelphia: W. B. Saunders Company, 1933.
3. Wartenberg R. The examination of reflexes, a simplification. Chicago: The yearbook publishers, 1945.
4. Burch GE. A Primer of Venous Pressure, Philadelphia: Lea & Febiger, 1950.
5. Walker H. Physical Diagnosis, St Louis: Mosby, 1952.
6. Fowler NO. Physical Diagnosis of Heart Disease, New York: Macmillan, 1962.
7. Martini P. Principles and Practice of Physical Diagnosis, 3rd edn, Philadelphia: Lippincott, 1962.
8. Perloff JK. Physical Examination of the Heart and Circulation, Philadelphia: W.B. Saunders, 1982.
9. Swartz MH. Textbook of Physical Diagnosis: History and Examination, 5th edn, Philadelphia: W.B. Saunders, 2006.
10. Evans LR, Bybee JR. Evaluation of student skills in physical diagnosis. J Med Educ 1965;40:199–204.
11. Wasson J, Sox HC, Jr, Tompkins RK, et al. Teaching physical diagnosis: The effect of a structured course taught by medical students. J Med Educ 1976;51(12):1014–5.
12. McGlynn TJ, Sayre A, Kennedy D. Physical diagnosis courses—A question of emphasis. J Fam Pract 1978;6(3):565–71.
13. Hunt DK, Badgett RG, Woodling AE, Pugh JA. Medical student career choice: Do physical diagnosis preceptors influence decisions? Am J Med Sci 1995;310(1):19–23.
14. Hamann C, Volkan K, Fishman MB, et al. How well do second-year students learn physical diagnosis? Observational study of an objective structured clinical examination (OSCE). BMC Med Educ 2002;2:1.
15. Fagan MJ, Griffith RA, Obbard L, O'Connor CJ. Improving the physical diagnosis skills of third-year medical students: A controlled trial of a literature-based curriculum. J Gen Intern Med 2003;18(8):652–5.
16. Nieman LZ, Cheng L, Hormann M, et al. The impact of preclinical preceptorships on learning the fundamentals of clinical medicine and physical diagnosis skills. Acad Med 2006;81(4):342–6.
17. Rosenbloom ST, Grande J, Geissbuhler A, Miller RA. Experience in implementing inpatient clinical note capture via a provider order entry system. J Am Med Inform Assoc 2004;11(4):310–5.
18. Shultz E, Rosenbloom T, Kiepek W, et al. Quill: A novel approach to structured reporting. AMIA Annu Symp Proc 2003;1074.
19. Bell DS, Greenes RA. Evaluation of UltraSTAR: Performance of a collaborative structured data entry system. Proc Annu Symp Comput Appl Med Care 1994:216–22.
20. Johnson KB, Cowan J, Clictate. A computer-based documentation tool for guideline-based care. J Med Syst 2002;26(1):47–60.
21. Kahn CE, Jr, Wang K, Bell DS. Structured entry of radiology reports using World Wide Web technology. RadioGraphics 1996;16(3):683–91.
22. E/M History Criteria. Family Practice Notebook. Available at: http://www.fpnotebook.com/MAN3.htm. Accessed Jul 13, 2007.
23. Leslie SJ, Hartswood M, Meurig C, et al. Clinical decision support software for management of chronic heart failure: Development and evaluation. Comput Biol Med 2006;36(5):495–506.
24. Kuperman GJ, Bobb A, Payne TH, et al. Medication-related clinical decision support in computerized provider order entry systems: A review. J Am Med Inform Assoc 2007;14(1):29–40.
25. Meystre S, Haug PJ. Automation of a problem list using natural language processing. BMC Med Inform Decis Mak 2005;5:30.
26. Spickard A, III, Gigante J, Stein G, Denny JC. Automatic capture of student notes to augment mentor feedback and student performance on patient write-ups. J Gen Intern Med 2008;23(7):979–84.
27. Spencer E, Swanson T, Hueston WJ, Edberg DL. Tools to improve documentation of smoking status. Continuous quality improvement and electronic medical records. Arch Fam Med 1999;8(1):18–22.
28. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12(4):448–57.
29. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402.
30. Jain NL, Knirsch CA, Friedman C, Hripcsak G. Identification of suspected tuberculosis patients based on natural language pro-

cessing of chest radiograph reports. Proc AMIA Annu Falls Symp 1996:542–6.
31. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Falls Symp 1997:829–33.
32. Xu H, Anderson K, Grann VR, Friedman C. Facilitating cancer research using natural language processing of pathology reports. Medinfo 2004;11(1):565–72.
33. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. J Biomed Inform 2006;39(6):589–99.
34. Hersh WR, Donohoe LC. SAPHIRE International: A Tool for Cross-Language Information Retrieval. Proc AMIA Symp 1998: 673–7.
35. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proc AMIA Symp 2001:17–21.
36. Denny JC, Smithers JD, Miller RA, Spickard A, III. "Understanding" medical school curriculum content using KnowledgeMap. J Am Med Inform Assoc 2003;10(4):351–62.
37. Denny JC, Spickard A, Miller RA, et al. Identifying UMLS concepts from ECG impressions using KnowledgeMap. AMIA Annu Symp Proc 2005:196–200.
38. Nadkarni P, Chen R, Brandt C. UMLS concept indexing for production databases: A feasibility study. J Am Med Inform Assoc 2001;8(1):80–91.
39. Elkin PL, Brown SH, Bauer BA, et al. A controlled trial of automated classification of negation from clinical notes. BMC Med Inform Decis Mak 2005;5(1):13.
40. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: Ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006;81(6): 741–8.
41. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5): 301–10.
42. Chapman WW, Cooper GF, Hanbury P, et al. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. J Am Med Inform Assoc 2003;10(5):494–503.
43. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform 2001;34(1):4–14.
44. Gupta D, Saul M, Gilbertson J. Evaluation of a deidentification (De-id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 2004;121(2):176–86.
45. Logical observation identifiers names and codes. Available at: http://www.regenstrief.org/medinformatics/loinc/. Accessed Jun 19, 2007.
46. Miller RA, Pople HE, Jr, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982;307(8):468–76.
47. LeMaire JB, Schaefer JP, Martin LA, et al. Effectiveness of the quick medical reference as a diagnostic tool. CMAJ 1999;161(6): 725–8.
48. Miller RA, McNeil MA, Challinor SM, Masarie FE, Jr, Myers JD. The INTERNIST-1/quick medical REFERENCE project—Status report. West J Med 1986;145(6):816–22.
49. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Info Med 1998;37(4–5):394–403.
50. Rosenbloom ST, Miller RA, Johnson KB, Elkin PL, Brown SH. Interface terminologies: Facilitating direct entry of clinical data into electronic health record systems. J Am Med Inform Assoc 2006;13(3):277–88.
51. Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO. The unified medical language system: An informatics research collaboration. J Am Med Inform Assoc 1998;5(1):1–11.
52. Crowell J, Zeng Q, Ngo L, Lacroix EM. A frequency-based technique to improve the spelling suggestion rank in medical queries. J Am Med Inform Assoc 2004;11(3):179–85.
53. Crowell JB, Zeng QT, Kogan S. A technique to improve the spelling suggestion rank in medical queries. AMIA Annu Symp Proc 2003:823.
54. Winawer S, Fletcher R, Rex D, et al. Colorectal cancer screening and surveillance: Clinical guidelines and rationale-update based on new evidence. Gastroenterol 2003;124(2):544–60.
55. Denny JC, Bastarache L, Sastre EA, Spickard A, III. Tracking medical students' clinical experiences using natural language processing. J Biomed Inform 2009.
56. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1(2):142–60.